Persistent Surveillance of Transient Events with Unknown Statistics

Cenk Baykal[1], Guy Rosman[1], Kyle Kotowick[1], Mark Donahue[2], and Daniela Rus[1]

***Abstract*— We consider the use of a mobile agent to monitor stochastic, transient events that occur in discrete locations in the environment with the objective of maximizing the number of event observations in a balanced manner. We assume that the events of interest at each station follow a stochastic process with an initially unknown and station-specific rate parameter. Consequently, the persistent monitoring problem we address in this paper is a bandit problem -similar to the canonical Multi-Armed Bandit problem- in which we are faced with the inherent trade-off between exploration and exploitation. We introduce a novel monitoring algorithm with provable guarantees that leverages variance estimates to generate policies capable of simultaneously taking into account the pertinent monitoring objectives and the balance between exploration and exploitation. We present analysis establishing lower bounds for the performance of our algorithm measured with respect to the quality of the policies generated. We present experimental results supporting our proposed algorithm and comparing its performance to that of current state-of-the-art monitoring algorithms.**

I. Introduction

We consider the problem of using a single mobile robot to monitor stochastic, transient events of interest occurring at discrete locations in the environment. We assume that events at each location follow a stochastic process with an unknown rate that is independent of other locations' rates. Since the events are stochastic and transient, their exact time of occurrence cannot be known apriori. Hence, the monitoring process requires the robot to visit each location and remain at that location for some amount of time in anticipation of events to occur.

An example of a surveillance task involving the monitoring of different bird species by a documentary maker is shown in Fig. 1. Additional examples of scenarios following this setting include robots patrolling the city in search of possible suspicious activities or mobile sensors roaming the environment to track wildlife around oases in the desert. The aforementioned scenarios each outline a persistent monitoring problem for which we would like to use a single mobile agent to monitor stochastic events in an information-driven way. The fact that we cannot concurrently monitor each location due to limited mobile resources motivates the need for optimal *monitoring policies*.

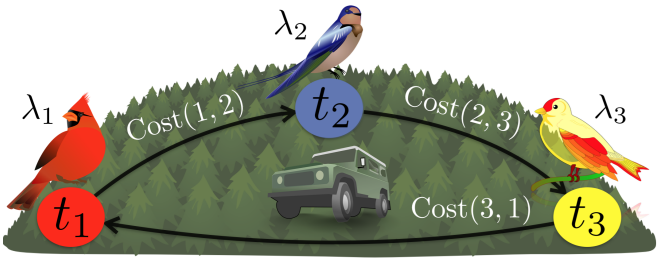

Fig. 1. A persistent monitoring application in which a documentary maker would like to monitor three different species of birds appearing in three discrete, species-specific locations. Bird sightings at each location follow a stochastic process with a rate that is initially unknown to the documentary maker and must be learned and approximated throughout the monitoring process. Given a cyclic path defining the sequence of stations to visit, the documentary maker would like to traverse this cyclic path repeatedly, stopping at each station for an appropriate amount of time to observe the birds.

We assume that we are given a cyclic patrolling route and seek to generate the optimal observation time to be spent at each location subject to a given optimality criteria [1], [2]. There may be several competing objectives of interest in a real-world monitoring scenario. These may include objectives pertaining to the number of events observed, the distribution of attention to all the stations, the time between consecutive observations at a station, and the classical trade-off between exploration and exploitation given the unknown rates of the stations. In this paper, we consider our overarching objective to be maximizing the number of observations across all stations in a balanced way while simultaneously balancing the inherent exploration and exploitation trade-off. We note this case can be extended to the case of reasoning over different trajectories as shown in [3].

Policy generation is rendered challenging by the fact that the exact timing of (stochastic) events cannot be predicted in advance and further the event statistics are assumed to be unknown apriori. These relaxed assumptions are in contrast to previous problem definitions such as those in [1], [2], [3], where the statistics of events occurring at different locations – such as rate of occurrence – were assumed to be known. In our case, the relaxation of this assumption results in the canonical exploration and exploitation problem, as the robot must simultaneously learn the statistics about events in the environment and adjust its policy in order to optimize the pertinent monitoring objective. The trade-off between exploration and exploitation that we address in this paper

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

[1]Cenk Baykal, Guy Rosman, Kyle Kotowick, and Daniela Rus are with the Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology (MIT), Cambridge, MA. {baykal,rosman,kotowick,rus}@csail.mit.edu

[2]Mark Donahue is with the MIT Lincoln Laboratory, Lexington, MA. mark.donahue@ll.mit.edu

is also faced by the canonical multi-armed bandit problem [4], [5] and reinforcement learning [6].

In this paper, we introduce a novel persistent monitoring algorithm with provable guarantees that quantifies and employs the uncertainty of our rate approximations to generate policies in order to reason about and explicitly consider the inherent exploration and exploitation trade-off. We present analysis proving probabilistic error bounds on the accuracy of rate approximations and the optimality of generated policies as a function of the number of the monitoring cycles. We present simulation results that compare the performance of our algorithm with that of an adaptive strategy and a state-of-the-art monitoring algorithm [2].

II. Related Work

In part due to the ubiquity of persistent monitoring tasks, the problem of persistent surveillance has been previously addressed with respect to a variety of applications and environments. For instance, in [7] the authors considered persistent surveillance of discrete locations -such as buildings, windows, doors- using a team of autonomous micro-aerial vehicles (MAVs). While UAVs are predominantly associated with persistent monitoring tasks, in [8] the authors considered the generation of monitoring policies for autonomous underwater vehicles with the objective of facilitating efficient high-value data collection. Furthermore, in [2] and [9], the authors present different approaches to the min-max latency walk problem in the context of discrete stations.

Persistent monitoring problems for multiple agents have also been studied with regard to a variety of different objectives. In [10], the authors consider the problem of controlling multiple agents to minimize an uncertainty metric in the context of a 1-D mission space. Furthermore, decentralized approaches to controlling a network of robots for purposes of sensory coverage has been investigated in [11], where the authors presented a control law to drive a network of mobile robots to an optimal sensing configuration.

In addition to persistent monitoring work in static environments, the case of dynamic environments has also been an avenue of interest [12], [13], [14]. Namely, authors of [12] considered optimal sensing in a time-changing Gaussian Random Field and proposed a new randomized path planning algorithm to find the optimal infinite horizon trajectory. In [13], the authors considered a changing environment modeled as a field which grows in locations that are not in the range of the robot and proposed a linear program to generate speed controllers capable of keeping the field bounded.

Persistent surveillance is inherently closely related to sensor scheduling [15], sensor positioning [16], and coverage problems [17]. Thus, previous approaches have considered the problem of persistent monitoring in the context of a mobile sensor [18]. For instance, in [19], the authors considered the problem of finding shortest *watchman routes* that enable the watchman to traverse paths along which every point in a given space is visible; the authors showed that this problem is NP-hard in general.

From the perspective of the persistent monitoring problem scenario that we address in this paper, our work can be seen as most similar to that of [2], where the authors considered the monitoring of stochastic, transient events occurring in discrete locations in the environment. The authors imposed the relatively strong assumption of having exact and full knowledge of the event statistics governing each stochastic process at each location prior to the monitoring process. In the context of this assumption, the authors presented a provably-optimal algorithm that generates the unique optimal policy maximizing the balance of observations while minimizing the maximum time between two consecutive observations at each station [2].

Viewed from the perspective of sequential decision making in the context of uncertainty, there exists parallels between the monitoring problem that we consider in this paper and the canonical problem of prediction with expert advice where the best expert is unknown apriori. An even more profound relationship and similarity exists between our problem and the widely-studied Multi-Armed Bandit (MAB) problem, in which a gambler is faced with a row of K slot machines that each yield a stochastic reward according to a machine-specific probability distribution with a finite mean which is initially unknown [4], [5]. The objective is to maximize accumulated reward by choosing the optimal machine to play at each discrete time step so that the expected regret with respect to the reward accumulated after a finite number of time is minimized.

There exist algorithms with provable regret guarantees even in the finite-horizon case for both the prediction for experts problem [20] and MAB [4], [5]. Unfortunately however, application or extension of these algorithms to the problem of persistent surveillance is rendered non-trivial due to salient differences between the persistent monitoring problem that we consider and a widely-studied bandit problem such as MAB. Namely, the persistent surveillance problem that we address in this paper exhibits a continuous state and parameter space, which is in contrast to MAB where the bandit attempts to choose the optimal lever to pull among a finite set of levers at discrete time steps, i.e. rounds. Furthermore, the monitoring problem we consider allows traveling the given cyclic path multiple times. This necessitates additional reasoning for iteration-dependent policies that consider the trade-off between the cost (i.e., wasted travel time that could otherwise be spent on observing) incurred by traveling from one station to the next and the total time that should be spent in traversing each monitoring cycle.

In contrast to all of the aforementioned prior work in the realm of persistent surveillance, we present algorithms with provable guarantees for the problem of monitoring of stochastic and transient events occurring in discrete stations in which the event statistics are unknown apriori. We employ Bayesian inference to efficiently learn, approximate, and reason about the event statistics at each station. Our algorithm explicitly quantifies and considers the uncertainties over our approximations to generate time-efficient, adaptive policies which simultaneously achieve near-optimal monitoring ob-

jective values and balance exploration and exploitation.

III. Problem Definition

Let there be $n \in \mathbb{N}_+$ stations, labeled by $i \in [n]$, whose locations are known. Events of interest occur at each station i and follow a Poisson process with an unknown rate parameter, denoted by λ_i, where the rate for each station is independent of other stations' rates. We assume that the stations are spatially distributed in the domain and hence the robot must spend a non-zero travel time $c_{i,j} \in \mathbb{R}_+$ as it travels from one station i to another station j.

We assume that we are given a cyclic path between the stations and our goal is to generate a policy stating the observation time that the robot should spend at each station to optimize an arbitrary monitoring objective. Over a monitoring period that is presumably bounded by resource constraints, a robot may traverse the given cyclic path multiple times and execute a variety of policies. We formally define a *monitoring cycle* as the complete execution of a monitoring policy and let $k \in \mathbb{N}_+$ denote each cycle. Under this terminology, a policy to be executed at cycle k, π_k, is defined as the sequence of observation times per station, i.e. $\pi_k := (t_{1,k}, t_{2,k}, \ldots, t_{n,k})$ where $t_{i,k} \in \mathbb{R}_+$ is the time to spend at each station $i \in [n]$.

Due to the presence of stochastic events, there will be inevitable variability from the execution of one monitoring cycle to the next. As will detailed further in Sec. IV, the number of events observed, $n_{i,k}$ and the time spent, $t_{i,k}$, at each station i constitute sufficient statistics to be considered at each cycle k; we let $X_i^k := (n_{i,k}, t_{i,k})$ denote this pair of values and the set of all relevant statistics obtained from the start of the monitoring task to iteration k as $X_i^{1:k} := \{X_i^1, X_i^2, \ldots, X^k\}$.

As mentioned in Sec. II, the majority of problems that face the exploration and exploitation trade-off, such as MAB, define the overarching optimization problem as the minimization of regret after a finite amount of time. MAB is concerned with the sole objective of maximizing the cumulative reward obtained, hence there is only one objective function (regarding the accumulated reward) that is considered. However, in the persistent monitoring problem that we address in this paper, the overarching goal is to *simultaneously* maximize the number of events observed and maximize the balance of observations across all stations. Consequently, the multi-objective problem we consider in persistent monitoring renders the definition of regret with respect to multiple objectives to be non-trivial.

We instead recast the problem of persistent monitoring in to an optimization problem with respect to individual monitoring cycles and present an alternative, *per-cycle* definition for the optimization problem. Defining the monitoring problem in this way that is local with respect to individual monitoring cycles can be viewed as a heuristic approach for greedily generating high-quality policies which perform well in minimizing regret with respect to both objectives.

We formalize our overarching objective functions as follows. We let $f_{\text{obs}}(\pi_k)$ be the objective function regarding the expected number of observations made across all stations, i.e.

$$f_{\text{obs}}(\pi_k) := \sum_{i=1}^{n} \mathbb{E}[N_i(\pi_k)] \tag{1}$$

where $\mathbb{E}[N_i(\pi_k)] = \lambda_i t_{i,k}$ by definition of expectation for a Poisson process with rate λ_i. In order to reason about balanced attention, we formalize the notion of observation balance by letting the function $f_{\text{bal}}(\pi_k)$ denote as in [2] the expected observations ratio for a given π_k which we seek to maximize,

$$f_{\text{bal}}(\pi_k) := \min_i \frac{\mathbb{E}[N_i(\pi_k)]}{\sum_{j=1}^{n} \mathbb{E}[N_j(\pi_k)]}. \tag{2}$$

The theoretical and idealized definition of persistent surveillance is traditionally defined as an infinite-horizon problem in which the total monitoring time is unbounded. Intuitively, we expect the agent execute multiple monitoring cycles of varying time length depending on iteration-specific policies that consider past history and observations. In light of a possibly unbounded monitoring time, the two aforementioned objective functions defined above do not help establish an appropriate upper bound on the total time that should be spent per monitoring cycle.

Rather than imposing an arbitrary bound on the monitoring time per cycle, we let the bound be a function of the uncertainty over the rates at each station. Namely, we seek to establish an adaptive bound on the observation time for each station in a way that considers the trade-off between travel-cost and the need to execute multiple monitoring cycles so that each station can be visited more than once. In what follows, we introduce a class policy optimization problems subject to the *uncertainty constraint*, a hard constraint that adaptively balances exploration and exploitation by controlling the decay of uncertainty over time. In short, the premise of the uncertainty constraint is to induce a rapid decrease of approximation uncertainty, which enables more accurate evaluations of prospective policies in the subsequent cycle, leading to the generation of high-quality policies within a short amount of time.

More formally, let $\upsilon_i : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be a function that quantifies the uncertainty in our estimate of the rate of each station i after a certain number of iterations. At each iteration k, having gathered and observed the events in the previous $k-1$ monitoring cycles, we would like to generate a policy π_k such that our uncertainty in our approximations decreases by some factor after executing π_k, with high probability. More formally, for a given $\delta \in (0,1), \varepsilon \in (0, \frac{1}{2})$, each policy π_k must satisfy the following uncertainty constraint $\forall i \in [n]$

$$\mathbb{P}\left(\upsilon_i(k|\pi_k) \leq \delta \upsilon_i(k-1) \middle| X_i^{1:k-1}\right) > 1 - \varepsilon. \tag{3}$$

In light of our monitoring objectives and the uncertainty constraint, we formalize the per-cycle optimization problem as follows.

Problem 1 (Per-cycle Monitoring Optimization Problem)**.** *In each iteration $k \in \mathbb{N}_+$ generate a policy π_k^* that simultaneously satisfies the uncertainty constraint (3) and maximizes*

the balance of observations, i.e.,

$$\pi_k^* \in \underset{\pi_k}{\operatorname{argmax}}\, f_{bal}(\pi_k) \tag{4}$$
$$s.t. \quad \mathbb{P}\left(\upsilon_i(k|\pi_k^*) \leq \delta\upsilon_i(k-1)\middle|X_i^{1:k-1}\right) > 1-\varepsilon \quad \forall i \in [n].$$

The per-cycle problem above defines the optimization to be solved at each cycle $k \in \mathbb{N}_+$ in order to generate an appropriate monitoring policy π_k. By associating the optimization problem with each cycle, we ensure that the generated policies are adaptive to the events that occur in previous cycles and enable the consideration of refined rate approximations. In the following section, we introduce a method for generating adaptive policies at each iteration that are optimal with respect to the optimization problem defined above.

IV. Methods

In this section, we introduce a novel monitoring algorithm that generates optimal policies at each monitoring cycle with respect to the optimization problem defined in Sec. III. We describe the sub-procedure for learning and approximating event statistics using Bayesian inference, which enables the incorporation of apriori knowledge and the generation of rate approximations for each station. We outline and provide pseudo-code for generating dynamic, adaptive policies that appropriately interleave learning and approximating event statistics (exploration) with generating and executing policies (exploitation).

A. Learning and Approximating Event Statistics

Prior to the monitoring process, we may have prior beliefs about what the rate λ_i could be for each station i. To model and incorporate any beforehand knowledge regarding the rate parameter, we use a Gamma distribution defined by the shape hyper-parameter α_i and the scale hyper-parameter β_i as the conjugate prior for the parameter λ_i. The hyper-parameters α_i and β_i will be initialized to values representing the prior beliefs, which we denote as the hyper-parameters $\alpha_{i,0}$ and $\beta_{i,0}$ and will then be updated during the monitoring process to represent our posterior beliefs given observations.

We can obtain the posterior distribution for the rate of any arbitrary station by updating the hyper-parameters α_i and β_i. Given the current values of α_i and β_i at cycle k, consider observing $n_{i,k}$ observations in $t_{i,k}$ time. Then, our posterior distribution in light of the observations $X^{1:k}$ is defined as:

$$\begin{aligned}\mathbb{P}(\lambda_i \mid X^{1:k}) &= \frac{\mathbb{P}(X_i^{1:k} \mid \lambda_i)\,\mathbb{P}(\lambda_i)}{\mathbb{P}(X_i^{1:k})} \\ &\propto \text{Gamma}(\alpha_i + n_{i,k}, \beta_i + t_{i,k}).\end{aligned} \tag{5}$$

where (5) follows by conjugacy. For any arbitrary number of $n_{i,k}$ events observed during $t_{i,k}$ time, the posterior update procedure simply entails updating the hyper-parameters based on their previous values and the values of $n_{i,k}$ and $t_{i,k}$, i.e. $\alpha_i \leftarrow \alpha_i + n_{i,k}$ and $\beta_i \leftarrow \beta_i + t_{i,k}$ at each cycle. More generally, after k monitoring cycles our posterior distribution is given by $\text{Gamma}(\alpha_{i,0} + \sum_{j=1}^{k} n_{i,j}, \beta_i + \sum_{j=1}^{k} t_{i,j})$.

After updating, we can employ the posterior distribution to generate a refined approximation, i.e. a point estimate, of the rate parameter for station i. Since our approximations will be iteratively changing, let $(\hat{\lambda}_i)_{k\in\mathbb{N}_+}$ denote the sequence of approximations for λ_i with respect to cycle k. For each cycle k we can leverage the fact that our updated posterior distribution is $\text{Gamma}(\alpha_{i,0} + \sum_{j=1}^{k} n_{i,j}, \beta_i + \sum_{j=1}^{k} t_{i,j})$ and set our approximation to be the posterior mean, i.e.,

$$\hat{\lambda}_{i,k} := E[\lambda_i|X^{1:k}] = \frac{\alpha_{i,0} + \sum_{i=1}^{n} n_{i,k}}{\beta_{i,0} + \sum_{i=1}^{n} t_{i,k}} = \frac{\alpha_i}{\beta_i}$$

which follows by definition of the Gamma distribution. Once updated approximations for the rates of all stations are available, i.e. $\hat{\lambda}_{1,k}, \ldots, \hat{\lambda}_{n,k}$, they can subsequently be used to evaluate the objective functions described in Sec. III.

B. Controlling Approximation Uncertainty

We formalize the definitions of the uncertainty function and the uncertainty constraint (3) introduced in Sec. III and present a method to generate policies subject to the uncertainty constraint. The premise of the uncertainty approach is to enable efficient generation of high-quality policies by enforcing a controlled and rapid expected decay of the uncertainty of our approximations with high probability.

Recall from Sec. III, at cycle k, $\upsilon_i : \mathbb{N}_+ \to \mathbb{R}_{\geq 0}$ is a function that quantifies our uncertainty in our rate estimate for the rate at station i. We choose to formally define the uncertainty function as the variance of the posterior distribution after k cycles, i.e.,

$$\upsilon_i(k) := Var(\lambda_i|X^{1:k}) = \frac{\alpha_i}{\beta_i^2} = \frac{\alpha_{i,0} + \sum_{i=1}^{n} n_{i,k}}{(\beta_{i,0} + \sum_{i=1}^{n} t_{i,k})^2}. \tag{6}$$

Under this setting, for a given policy π_k at cycle $k \in \mathbb{N}_+$ the uncertainty constraint as defined in Sec. III is equivalent to:

$$\mathbb{P}\left(Var(\lambda_i|X^{1:k}) \leq \delta Var(\lambda_i|X_i^{1:k-1})|X_i^{1:k-1}\right) > 1-\varepsilon$$

for all stations $i \in [n]$. We further simplify the uncertainty constraint by employing the definition of posterior variance and obtain

$$\mathbb{P}\left(\frac{\alpha_i + N_{i,k}(\pi_k)}{(\beta_i + t_{i,k})^2} \leq \delta\frac{\alpha_i}{\beta_i^2}|X_i^{1:k-1}\right) \tag{7}$$
$$= \mathbb{P}\left(N_{i,k}(t_{i,k}) \leq \delta K(t_{i,k})|X_i^{1:k-1}\right) > 1-\varepsilon \tag{8}$$

where $N_{i,k}(t_{i,k}) \sim \text{Poisson}(\lambda_i t_{i,k})$ and $K(t_{i,k}) := \delta\frac{\alpha_i}{\beta_i^2}(\beta_i + t_{i,k})^2 - \alpha_i$.

Generating an appropriate $t_{i,k}$ that satisfies the inequality given by (8) above requires that we reason about the possible values that the random variable $N_{i,k}(\pi_k)$ can assume. Hence, in order to make a more informed decision in generating the observation time $t_{i,k}$, we leverage the notion of a credible interval in our policy generation process. Namely, given a fixed $\varepsilon \in (0, \frac{1}{2})$ and past observations $X_i^{1:k-1}$ at station i, we construct a credible interval for λ_i denoted by the open set $C_i(X_i^{1:k-1}) := (\lambda_i^{\text{l}}, \lambda_i^{\text{u}})$ such that:

$$\forall\lambda_i \in \mathbb{R}_+\, \mathbb{P}(\lambda_i \in (\lambda_i^{\text{l}}, \lambda_i^{\text{u}})|X_i^{1:k-1}) = 1-\varepsilon$$

for the rate parameter λ_i of each station i. In constructing the the credible interval, we reason about the regularized Gamma function, denoted by $Q(a,s)$, due to its inherent relationship with the cumulative distribution function of the posterior Gamma distribution. Namely, we employ the inverse of the regularized Gamma function with respect to the second variable, $Q^{-1}(a,s)$, to generate an equal-tailed credible interval $C_i(X_i^{1:k-1})$ as follows:

$$\lambda_i^{\mathrm{l}} := \frac{Q^{-1}(\alpha_i, 1-\frac{\varepsilon}{2})}{\beta_i} \quad \lambda_i^{\mathrm{u}} := \frac{Q^{-1}(\beta_i, \frac{\varepsilon}{2})}{\beta_i}.$$

In addition, we have the property that

Since we have constructed an equal-tails credible interval with respect to $\varepsilon \in (0, \frac{1}{2})$, the following holds by definition:

$$\forall \lambda_i \in \mathbb{R}_+ \; \mathbb{P}(\lambda_i^{\mathrm{l}} > \lambda_i | X_i^{1:k-1}) = \mathbb{P}(\lambda_i^{\mathrm{u}} < \lambda_i | X_i^{1:k-1}) = \frac{\varepsilon}{2}.$$

Now, putting it all together, given observations $X_i^{1:k-1}$ after having executed $k-1$ cycles and the end points of the confidence interval λ_i^{l} and λ_i^{u}, generating an optimal observation time $t_{i,k}$ for each station i entails efficiently generating an observation time $t_{i,k}^*$ that satisfies the inequality given by (8). A notable observation in this context is the existence of an uncountably infinite values of $t_{i,k}$ that satisfy this inequality. This follows from the fact that the expression $K(t_{i,k})$ follows a quadratic relationship with respect to $t_{i,k}$, whereas a linear relationship exists in the expression defining the distribution's parameter, i.e. $\lambda_i t_{i,k}$. Recalling that achieving low-latency, i.e. monitoring cycles with small amount of observation times, is preferable for a monitoring task, we pick the minimum $t_{i,k}$ possible.

In other words, the optimal $t_{i,k}^*$ is defined as the optimal solution to the following optimization:

$$\begin{aligned} &\inf_{t_{i,k} \in \mathbb{R}_+} t_{i,k} \\ &\text{s.t.} \quad \mathbb{P}\left(N_{i,k}(t_{i,k}) \leq \delta K(t_{i,k}) | X_i^{1:k-1}\right) > 1-\varepsilon. \end{aligned}$$

Unfortunately, due to the inherent complexity of the cumulative distribution function for the Poisson random variable, using a non-linear optimization method to generate the optimal solution is rendered computationally expensive. Hence, instead of performing exact computation for the quantile function, we employ a sharp inequality provided by [21] which improves upon the Chernoff-Hoeffding inequalities by a factor of at least two in approximating the cumulative distribution function introduced.

As demonstrated rigorously in the analysis section (Sec. V), the use of this approximation combined with the observation regarding the quadratic-linear relationship of $K(t_{i,k})$ and $\lambda_i t_{i,k}$ results in a simplified solution for $t_{i,k}^*$. Namely, an appropriate choice of $t_{i,k}^*$ is given by

$$t_{i,k}^* := t \in \mathbb{R}_+ \mid H(\lambda_i^{\mathrm{u}} t, K(t)) - \frac{1}{2} W\left(\frac{(\varepsilon-2)^2}{2\varepsilon^2 \pi}\right) = 0, \quad (9)$$

where $H(m,k)$ is the Kullback-Leibler (KL) divergence between two Poisson distributed random variables with means m and k and W is the Lambert W function. . An appropriate value for $t_{i,k}^*$ can be efficiently obtained by invoking a root-finding algorithm such as Brent's method on equation above.

The constant factor $\delta \in (0,1)$ controls the rapidness of uncertainty decay. There exists a tradeoff between low values of δ, which lead to lengthy, but also risky policies, and high values for δ which lead to shorter, but less efficient policies due to incurred travel time. Thus, in order to pick an appropriate value of δ, we use the following generalized sigmoid function that takes into account the number of stations and the total travel time per cycle:

$$\delta(n) := \delta_{\min} + \frac{\delta_{\max} - \delta_{\min}}{1 + e^{-T_{\mathrm{tr}} n}}$$

where $\delta_{\min}, \delta_{\max} \in (0,1), T_{\mathrm{tr}} := (\sum_{i=1}^{n-1} c_{i,i+1} + c_{n,1})^{-1} \in \mathbb{R}_+$ are the lower asymptote, the upper asymptote, and the growth rate respectively. where $\delta_{\min} = 1/4, \delta_{\max} = 0.99, \delta_{\mathrm{r}} = (\sum_{i=1}^{n-1} c_{i,i+1} + c_{n,1})^{-1}$ are the lower asymptote, the upper asymptote, and the growth rate respectively. It is worth noting that δ is a static variable and is initialized only once in the beginning of the monitoring process.

C. Generating Balanced Policies that Consider Approximation Uncertainty

We extend the method defined in the previous section so that the generated policy π_k^* simultaneously satisfies the uncertainty constraint and balances attention given to all stations in the minimum time possible. The key insight behind our approach is that if we first compute a $\pi_{\mathrm{low}} := (t_1^{\mathrm{low}}, \ldots, t_n^{\mathrm{low}})$ where each t_i^{low} is defined by the expression given by Eq. (9) acts as a lower bound on each of the observation times. In other words, any observation time t_i for a particular station i that is higher than t_i^{low} given by π_{low} is ensured to satisfy uncertainty constraint by monotonicity as described in Sec. V. Now, we can initially set $\pi_k := \pi_{\mathrm{low}}$ to ensure that π_k and any policy with higher observation times satisfies the uncertainty constraint.

In addition to satisfying the uncertainty constraint on the observation times, we must also satisfy the balance constraint, i.e. maximize objective function 2. The idea is to generate a new policy π_k^* by increasing the observation times of π_k so that π_k^* satisfies the balance and uncertainty constraints. Note that π_k^* achieves the optimal balance value if and only if:

$$\mathbb{E}[N_1(\pi_k^*)] = \mathbb{E}[N_2(\pi_k^*)] = \cdots = \mathbb{E}[N_n(\pi_k^*)] = \hat{\lambda}_{n,k} t_{n,k}^*.$$

For the initial lower bound policy $\pi_k = \pi_{\mathrm{low}} = (t_1^{\mathrm{low}}, \ldots, t_n^{\mathrm{low}})$ from the expression in the previous section, it may very well be the case that the above equality does not hold. However, π_k can be modified by first looking at the maximum number of expected events that needs to be matched, i.e. $N_{\max} := \max_{i \in [n]} \hat{\lambda}_i t_i^{\mathrm{low}}$.

Now, using $N_{\max}$ we can increase the observation times in π_k to generate the true optimal policy $\pi_k^* = (t_{1,k}^*, \ldots, t_{n,k}^*)$. The generation procedure for each observation time is as follows:

$$t_{i,k}^* := \frac{N_{\max}}{\hat{\lambda}_{i,k}} = N_{\max} \frac{\beta_i}{\alpha_i}.$$

Our policy generating function that summarizes the sequence of steps described above is shown in Alg. 2. The entirety of our monitoring algorithm which employs Alg. 2 as a subprocedure to generate policies is shown as Alg. 1.

Algorithm 1 Executes the entirety of the monitoring process and employs Algorithm 2 as a sub-procedure to generate policies.

Input:

$(\alpha_{i,0}, \beta_{i,0})$: prior parameters for each station i
$c_{i,j}$: travel times for all pairs of stations i, j

$\alpha_i \leftarrow \alpha_{i,0}$
$\beta_i \leftarrow \beta_{i,0}$
$T_{\text{tr}} \leftarrow \sum_{i=1}^{n-1} c_{i,i+1} + c_{n,0}$
while NotDoneMonitoring() **do**
  for $i \in [n]$ **do**
    $\hat{\lambda}_i \leftarrow \frac{\alpha_i}{\beta_i}$
  $\pi^* \leftarrow \text{Algorithm2}(\alpha_i, \beta_i, T_{\text{tr}})$
  // Monitoring loop
  for $i \in [n]$ **do**
    // Observe events for time t_i^*
    $N_{\text{observed}} \leftarrow \text{ObserveEvents}(t_i^*)$
    // Update hyper-parameters of the posterior
    $\alpha_i \leftarrow \alpha_i + N_{\text{observed}}$
    $\beta_i \leftarrow \beta_i + t_i^*$
return $\pi^* = (t_1^*, \ldots, t_n^*)$

V. Analysis

In this section, we present analysis proving the fact that the each iterations-specific policy generated by our algorithm described in Sec. IV is an optimal solution with respect to the optimization problem defined in Sec. III with respect to the generated rate approximations $\hat{\lambda}_{i,k}, i \in [n]$ at each iteration $k \in \mathbb{N}_+$. Subsequently, we establish guarantees on the posterior variance and absolute error of our rate approximations as a function of optimization iterations. We conclude by employing the aforementioned properties to establish a bound on the quality of our generated solutions with respect to those generated by an *Oracle Algorithm* that is assumed to have perfect knowledge of the ground-truth rates.

We begin by showing that at every monitoring iteration $k \in \mathbb{N}_+$, the policy defined by the sequence of observation times, each generated according to the expression in 9 presented in Sec. IV, is optimal with respect to the rate approximations.

Lemma 1 (Satisfaction of the Uncertainty Constraint)**.** *For a given $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$ at iteration $k \in \mathbb{N}$, a value of $t_{i,k}^*$ given by Eq. (9) satisfies the uncertainty constraint (Eq. 3).*

Lemma 2 (Optimality of Generated Observation Times)**.** *For all $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$ at iteration $k \in \mathbb{N}$, an optimal observation time for each station with respect to optimization problem 1 presented in Sec. III is given by*

$$t_{i,k}^* := \frac{N_{max}}{\hat{\lambda}_{i,k}} = N_{max} \frac{\beta_i}{\alpha_i}$$

where $N_{max} := \max_{i \in [n]} \hat{\lambda}_{i,k} t_{i,k}^{low}$ and $t_{i,k}^{low}$ is given by expression (9).

Algorithm 2 Generates to an optimal policy with respect to optimization problem 1 defined in Sec. III

Input:

(α_i, β_i): hyper-parameters of the posterior
T_{tr}: total travel time for the given cycle
$\varepsilon \in (0, \frac{1}{2})$: user-given input

Output:

$\pi^* = (t_1^*, \ldots, t_n^*)$: optimal policy

// Calculate the lower bound for π
$\delta \leftarrow \delta_{\min} + \frac{\delta_{\max} - \delta_{\min}}{1 + e^{-n T_{\text{tr}}}}$
// Upper end-point of the $1-\varepsilon$ credible interval
$\lambda_{\text{u}} \leftarrow \frac{Q^{-1}(\beta_i, \frac{\varepsilon}{2})}{\beta_i}$
for $i \in [n]$ **do**
  // Define the function $K_i(t)$
  $K_i(t) := \delta \frac{\alpha_i}{\beta_i^2} (\beta_i + t_{i,k})^2 - \alpha_i$
  $t_i^{\text{low}} \leftarrow t \in \mathbb{R}_+ \mid H(\lambda_{\text{u}} t, K_i(t)) - \frac{1}{2} W\left(\frac{(\varepsilon-2)^2}{2\varepsilon^2 \pi}\right) = 0$
$\pi_{\text{low}} \leftarrow (t_1^{\text{low}}, \ldots, t_n^{\text{low}})$
// Calculate the maximum $\mathbb{E}[N_i(\pi_{\text{low}})]$
$N_{\max} \leftarrow \max_{i \in [n]} t_i^{\text{low}} \frac{\alpha_i}{\beta_i}$
// Balance attention based on $N_{\max}$
for $i \in [n]$ **do**
  $t_i^* \leftarrow N_{\max} \frac{\beta_i}{\alpha_i}$
return $\pi^* = (t_1^*, \ldots, t_n^*)$

In light of the appropriateness of our choices for the observation time, we can establish further guarantees that pertain to the posterior variance and the error of our approximations.

Lemma 3 (Bound on Posterior Variance)**.** *For any $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, the posterior variance $Var(\lambda_i | X^{(1:k)})$ is bounded above by $\delta^k Var(\lambda_i)$ with probability at least $(1-\varepsilon)^k$, i.e.,*

$$\mathbb{P}\left(Var(\lambda_i | X_i^{(1:k)}) \leq \delta^k Var(\lambda_i) | X^{(1:k)}\right) > (1-\varepsilon)^k$$

for all stations $i \in [n]$ where $Var(\lambda_i) := \frac{\alpha_{i,0}}{\beta_{i,0}^2}$ is the prior variance.

Corollary 4 (Bound on Approximation Variance)**.** *For any $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, the variance of our approximation $Var\left(\hat{\lambda}_{i,k} | X^{(1:k-1)}\right)$ is bounded above by $\delta^{k-1} Var(\lambda_i)$ with probability greater than $(1-\varepsilon)^{k-1}$, i.e.,*

$$\mathbb{P}\left(Var(\hat{\lambda}_{i,k} | X^{(1:k-1)}) \leq \delta^{k-1} Var(\lambda_i) | X^{(1:k-1)}\right) > (1-\varepsilon)^{k-1}$$

for all stations $i \in [n]$.

Theorem 5 (ξ-Bound on the Approximation Error)**.** *For any $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, for any $\xi \in \mathbb{R}_+$, our approximation $\hat{\lambda}_{i,k}$ lies within a ball of radius ξ centered at λ_i with probability at least $(1-\varepsilon)^{k-1}(1-\frac{\delta^{k-1}Var(\lambda_i)}{\xi^2})$, i.e.,*

$$\mathbb{P}\left(|\hat{\lambda}_{i,k} - \lambda_i| < \xi | X^{(1:k-1)}\right) > (1-\varepsilon)^{k-1}\left(1 - \frac{\delta^{k-1}Var(\lambda_i)}{\xi^2}\right)$$

for all $i \in [n]$.

Theorem 6 (Δ-Bound on Policy Optimality)**.** *For any $\xi_i \in \mathbb{R}_+$, $i \in [n]$, given that $0 < |\hat{\lambda}_{i,k} - \lambda_i| < \xi_i$ with probability as given in Theorem 5, let $\sigma_{min} := \sum_{i=1}^{n}(\lambda_i - \xi_i)^{-1}$ and $\sigma_{max} := \sum_{i=1}^{n}(\lambda_i + \xi_i)^{-1}$. Then, the objective value of our policy π_k^* at iteration k is within a factor of Δ of the ground-truth optimal solution, where $\Delta := \frac{\sigma_{min}}{\sigma_{max}}$ with probability greater than $(1-\varepsilon)^{n(k-1)}\left(1 - \frac{\delta^{k-1}Var(\lambda_i)}{\xi^2}\right)^n$.*

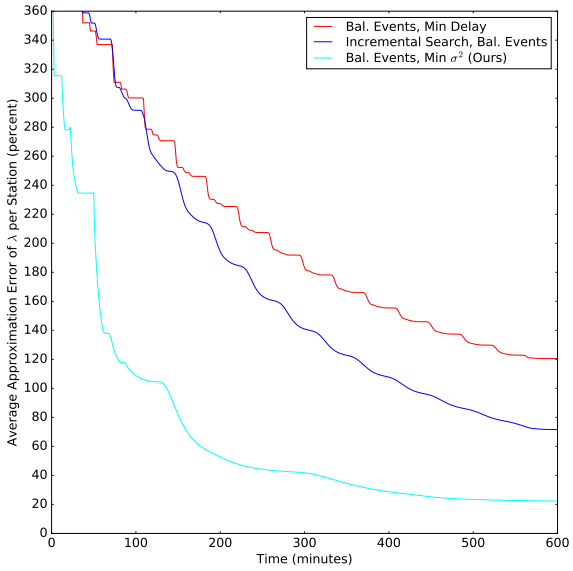

Fig. 2. We present results conveying the quality of the statistics approximations as a function of monitoring time. The rapid rate of approximation error for the performance of our algorithm (cyan) supports the conjecture that our algorithm is able to generate adaptive policies conducive to an accelerated rate of error decrease in contrast to the other algorithms' performance.

VI. Results

In this section, we present results that portray the performance of our algorithm in a monitoring scenario and contrast the quality of policies generated by our algorithm to that of a current state-of-the-art algorithm for persistent surveillance [2] and a dynamic algorithm that represents a naive method of generating adaptive policies. We consider the results of the experiments in two settings: (i) a synthetic simulation scenario in which the events are generated according to a Poisson distribution and (ii) a real-world inspired scenario

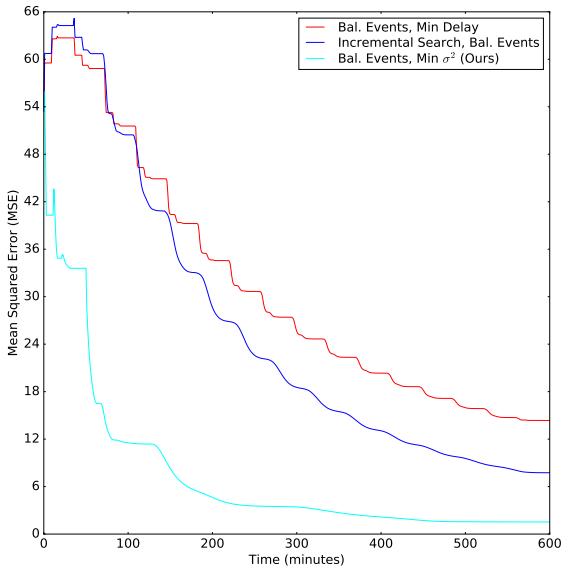

Fig. 3. Additional results of the performance of our algorithm with respect to the Mean Squared Error (MSE) metric as a function of monitoring time. Similar to the results shown in Fig. 2, we note that our algorithm (cyan) is able to obtain accurate approximations for event statistics quickly, resulting in smaller values for MSE in when compared to those of the other monitoring algorithms' values (red and blue).

-denoted as the *yellow backpack scenario*- simulated in the ARMA, a tactical military game.

The synthetic simulation framework and the monitoring algorithm were implemented in Python. The experiments were conducted on a MacBook Pro with one 3.1 GHz Intel Core i7 (4 cores total) processor and 16 GB of RAM. In what follows, we present the experimental scenarios and the respective results.

A. Synthetic Simulation Results

We obtained results from 10,000 trials (per algorithm) of a simulated persistent monitoring scenario involving the monitoring of events in 3 discrete stations for a monitoring period of 10 hours (600 minutes). The settings for the environment and the ground-truth rates were randomly generated by generating random variables from the following distributions for each of the three stations:

1) Prior Hyper-parameter $\alpha_{i,0} \sim \text{Uniform}(1,20)$
2) Prior Hyper-parameter $\beta_{i,0} \sim \text{Uniform}(0.5,1)$
3) Rate parameter $\lambda_i \sim \text{Uniform}(\frac{\alpha_{i,0}}{4\beta_{i,0}}, \frac{4\alpha_{i,0}}{\beta_{i,0}})$ events per
4) Cost of travel from station i to an adjacent station j $c_{i,j} \sim \text{Uniform}(2,5)$ minutes of travel time.

In this synthetic simulation, the arrival times of the random events were specifically drawn from a Poisson distribution

To ensure consistency and compare the algorithms in a fair manner, we incorporated the same learning and approximation procedure detailed in Sec. IV for the other two algorithms. This integration enabled us to measure the

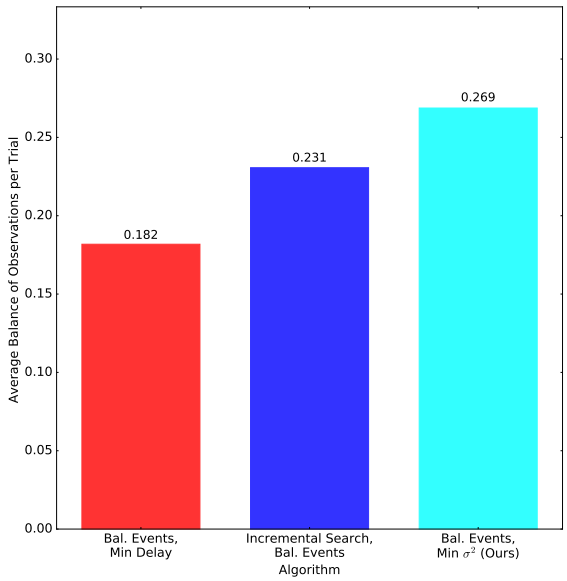

Fig. 4. The performance of our monitoring algorithms with respect to the objective function pertaining to the balance of observations. We can see that when the balance of observations is considered with respect to the entire 10 hour monitoring window, the policies generated by our algorithm achieve a significantly higher objective value than do the those generated by the other two algorithms.

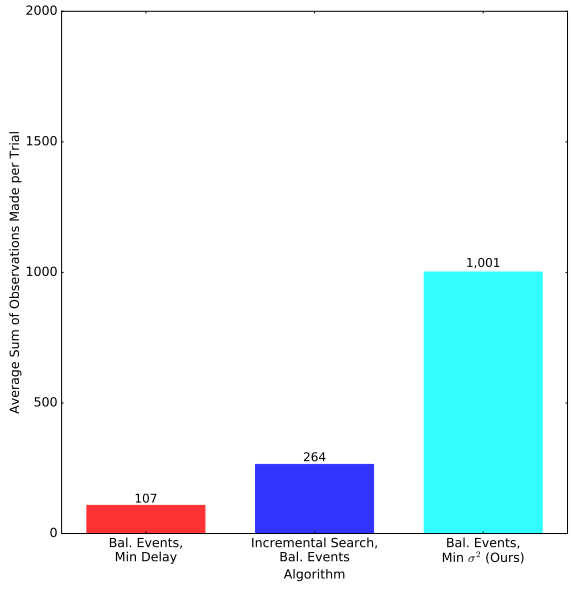

Fig. 5. The performance of each algorithm with respect to the objective pertaining to the number of events observed within the allotted monitoring time (10 hours). We note that our algorithm (cyan) enables the agent to observe significantly more events.

performance of an algorithm by that operated under the assumption of known rates prior to the monitoring procedure [2].

The label and description of each algorithm along with its corresponding color in the figures are as follows:

1) Bal. Events, Min. Delay (Red): the algorithm introduced by [2] which, as mentioned in Sec. II, assumes that the event statistics are available apriori.
2) Incremental Search, Bal. Events (Dark Blue): an algorithm that acknowledges the presence of the exploration/exploitation trade-off and attempts to generate adaptive and lengthier policies. The algorithm initially begins with a random upper bound on the total cycle time. After each monitoring iteration, the algorithm increases the upper bound monotonically by a small random amount (with an expected increase of 5 minutes) by generating observation times that balance expected observations subject to an arbitrary upper bound on the total cycle time.
3) Bal. Events, Min σ^2 (Cyan): our algorithm introduced in this paper that employs variance estimates to simultaneously generate policies and balance the exploration/exploitation trade-off in a near-optimal way.

We show plots of relative approximation error as a function of time, the total number of events observed on average after 10 hours of monitoring, the balance of observations with respect to all of the observations made in the 10 hour monitoring period, and the total computation time spent for generating policies during the execution of a trial. As expected, the results show that our algorithm shown in cyan in all figures is able to relatively outperform the other two evaluated algorithms with respect to every metric. Namely, from the figures we can see that our algorithm is able to efficiently generate balanced policies leading to policies capable of achieving near-optimal monitoring objective values while simultaneously inducing a rapid decline of approximation uncertainty.

B. ARMA Simulation Results

To be completed: in this subsection, we present the results regarding the yellow backpack scenario: a real-world inspired monitoring application involving the surveillance of people wearing yellow backpacks in an ARMA simulation (see Figs. 8 and 9).

VII. Conclusion

In this paper we introduced novel algorithms and objective criteria for the task of persistent monitoring of events with statistics that are unknown a priori. Our algorithms bridged previous literature and tools pertaining persistent surveillance and machine learning in order to introduce algorithms that were able to simultaneously explore and exploit the environment with respect to a given monitoring objective. Namely, our algorithms considered maximizing the number of observations across all stations in a balanced manner while simultaneously ensuring the controlled decay of uncertainty in our rate approximations . We presented analysis showing the favorable properties of our algorithm with regard to

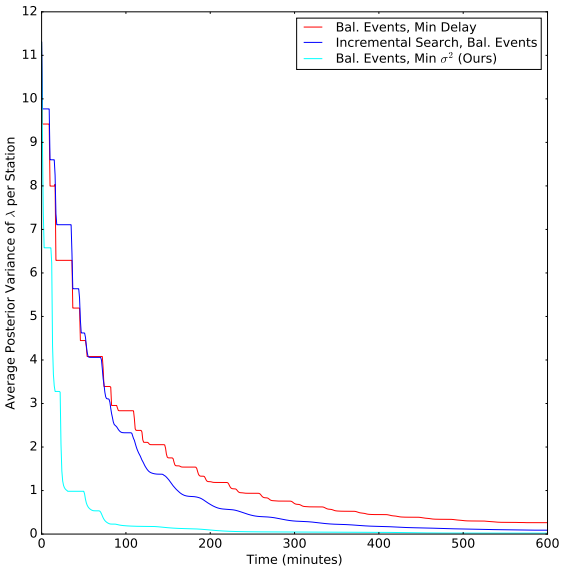

Fig. 6. We present results conveying the decaying behavior of the posterior variance as a function of time in a scenario involving 3 stations. We note that while the variance is monotonically decreasing in expectation for all algorithms compared, our algorithm (cyan) enables a much more rapid decay of the variance.

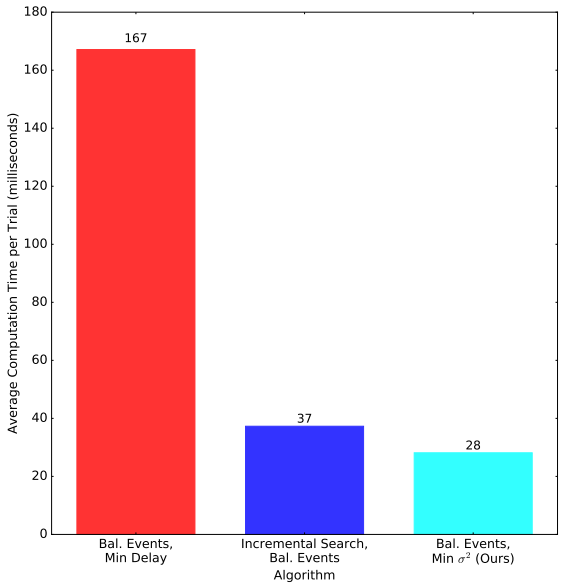

Fig. 7. We present results showing the computation time required to generate the policies in the simulated scenario. We can see that our algorithm (cyan) is able to generate high-quality policies with higher computational efficiency when compared to the other state-of-the-art algorithms.

uncertainty and policy optimality. We performed computational experiments with a diverse environment in terms of

Fig. 8. **Work in progress:** a rendition of the yellow backpack scenario simulated in ARMA. In this scenario, agents randomly wander around a town-like environment that includes buildings and apartments. The persistent monitoring task for a robot -such as a UAV- is to continuously survey the environment by following a given circular patrolling cycle, observing for an appropriate amount of time at particular locations, and detecting potentially malicious people wearing yellow backpacks during the observation process.

Fig. 9. **Work in progress:** a viewpoint of the yellow backpack simulation in ARMA different from that of Fig. 8 is shown. Three -potentially suspicious- agents carrying yellow backpacks are seen near a building located at the intersection of two streets.

event statistics and compared our monitoring approach to the state-of-the-art. In future work we intend to relax the assumptions imposed on the events further and extend our work to dynamic and large-scale environments.

REFERENCES

[1] Mac Schwager, Daniela Rus, and Jean-Jacques E. Slotine. Decentralized, adaptive coverage control for networked robots. *IJRR*, 28(3):357–375, 2009.

[2] J. Yu, S. Karaman, and D. Rus. Persistent monitoring of events with stochastic arrivals at multiple stations. *IEEE Transactions on Robotics*, 31(3):521–535, 2015.

[3] Daniel E. Soltero, Mac Schwager, and Daniela Rus. Generating informative paths for persistent sensing in unknown environments. In *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots Systems (IROS)*, pages 2172–2179, 2012.

[4] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[5] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[6] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, pages 237–285, 1996.

[7] Nathan Michael, Ethan Stump, and Kartik Mohta. Persistent surveillance with a team of mavs. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.

[8] Ryan N Smith, Mac Schwager, Stephen L Smith, Burton H Jones, Daniela Rus, and Gaurav S Sukhatme. Persistent ocean monitoring with underwater gliders: Adapting sampling resolution. *Journal of Field Robotics*, 28(5):714–741, 2011.

[9] Soroush Alamdari, Elaheh Fata, and Stephen L Smith. Persistent monitoring in discrete environments: Minimizing the maximum weighted latency between observations. *The International Journal of Robotics Research*, 33(1):138–154, 2014.

[10] Christos Cassandras, Xuchao Lin, and Xuchu Ding. An optimal control approach to the multi-agent persistent monitoring problem. *Automatic Control, IEEE Transactions on*, 58(4):947–961, 2013.

[11] Mac Schwager, Daniela Rus, and Jean-Jacques Slotine. Decentralized, adaptive coverage control for networked robots. *The International Journal of Robotics Research*, 28(3):357–375, 2009.

[12] Xiaodong Lan and Mac Schwager. Planning periodic persistent monitoring trajectories for sensing robots in gaussian random fields. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2415–2420. IEEE, 2013.

[13] Stephen L Smith, Mac Schwager, and Daniela Rus. Persistent robotic tasks: Monitoring and sweeping in changing environments. *Robotics, IEEE Transactions on*, 28(2):410–426, 2012.

[14] Daniel E Soltero, Mac Schwager, and Daniela Rus. Generating informative paths for persistent sensing in unknown environments. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2172–2179. IEEE, 2012.

[15] Ying He and Edwin KP Chong. Sensor scheduling for target tracking in sensor networks. In *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, volume 1, pages 743–748. IEEE, 2004.

[16] Alfred O Hero III, Christopher M Kreucher, and Doron Blatt. Information theoretic approaches to sensor management. In *Foundations and applications of sensor management*, pages 33–57. Springer, 2008.

[17] Yoav Gabriely and Elon Rimon. Competitive on-line coverage of grid environments by a mobile robot. *Computational Geometry*, 24(3):197–224, 2003.

[18] Jerome Le Ny, Munther A Dahleh, Eric Feron, and Emilio Frazzoli. Continuous path planning for a data harvesting mobile server. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 1489–1494. IEEE, 2008.

[19] Wei-pang Chin and Simeon Ntafos. Optimum watchman routes. *Information Processing Letters*, 28(1):39–44, 1988.

[20] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[21] Michael Short. Improved inequalities for the poisson and binomial distribution and upper tail quantile functions. *ISRN Probability and Statistics*, 2013, 2013.

VIII. Appendix

A. Proofs of Results Presented in Sec. V

1) Proof of Lemma 1:

Lemma 1 (Satisfaction of the Uncertainty Constraint)**.** *For a given $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$ at iteration $k \in \mathbb{N}$, a value of $t^*_{i,k}$ given by Eq. (9) satisfies the uncertainty constraint (Eq. 3).*

Proof. We first show that the proposed value of $t^*_{i,}$ satisfies the uncertainty condition (3) for all stations $i \in [n]$. Recall from Sec. IV that the uncertainty constraint is equivalent to the following:

$$\mathbb{P}\left(N_{i,k}(t^*_{i,k}) \leq \delta K(t^*_{i,k}) | X_i^{(1:k-1)}\right) > 1-\varepsilon \tag{10}$$

with $N_i(t^*_{i,k}) \sim \text{Poisson}(\lambda_i t^*_{i,k})$ and $K(t^*_{i,k}) := \delta \frac{\alpha_i}{\beta_i^2}(\beta_i + t^*_{i,k})^2 - \alpha_i$. Now, we can employ the credible interval established in Alg. 2 to further simplify the left-hand side of (10):

$$\begin{aligned}
&\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}\right) \qquad (11)\\
&= \int_0^{\infty} \mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda\right) \mathbb{P}\left(\lambda | X_i^{(1:k-1)}\right) d\lambda \\
&> \int_0^{\lambda_u} \mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda\right) \mathbb{P}\left(\lambda | X_i^{(1:k-1)}\right) d\lambda \\
&\geq \mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_u\right) \int_0^{\lambda_u} \mathbb{P}\left(\lambda | X_i^{(1:k-1)}\right) d\lambda \\
&= (1-\frac{\varepsilon}{2}) \mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_u\right) \qquad (12)
\end{aligned}$$

where we utilized the generated credible interval for λ_i and the fact that

$$\begin{aligned}
&\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_u\right) \\
&\quad = \inf_{\lambda \in (0, \lambda_u)} \mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda\right)
\end{aligned}$$

to establish the inequalities.

We can further simplify the expression in (12) by establishing a lower bound for the cumulative distribution function of a Poisson random variable with mean $m(t^*_{i,k}) = \lambda_u(x_i) t^*_{i,k}$, given the value $K(t^*_{i,k})$. Using the inequality established by [21], we have that the following holds for $k \geq m$:

$$\mathbb{P}\left(N_i(t^*_{i,k}) \leq k\right) > 1 - \frac{e^{-H(m,k)}}{\max\left\{2, \sqrt{4\pi H(m,k)}\right\}} \tag{13}$$

where $m := \mathbb{E}[N_i(t^*_{i,k}) | \lambda_u] = \lambda_u t^*_{i,k}$, $k = K^*_{i,k}$, and $H(m,k)$ is the Kullback-Leibler (KL) divergence between two Poisson distributed random variables with means m and k defined as

$$H(m,k) := m - k + k \ln\left(\frac{k}{m}\right).$$

We note that by definition of $t^*_{i,k}$, $H(m(t^*_{i,k}), K(t^*_{i,k})) = H^* = W\left(\frac{(\varepsilon-2)^2}{2\varepsilon^2\pi}\right)$, we have:

$$1 - \frac{e^{-H^*}}{\max\left\{2, \sqrt{4\pi H^*}\right\}} = 1 - \frac{\varepsilon}{2-\varepsilon},$$

and thus

$$\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_u\right) > 1 - \frac{\varepsilon}{2-\varepsilon}.$$

Continuing from (12) in light of this inequality yields

$$\begin{aligned}
&\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}\right) \\
&> (1-\frac{\varepsilon}{2}) \mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}, \lambda_u\right) \\
&> (1-\frac{\varepsilon}{2})(1-\frac{\varepsilon}{2-\varepsilon}) = 1-\varepsilon.
\end{aligned}$$

Putting it all together, we have for our choice of $t^*_{i,k}$ given by (9) that for any $\varepsilon \in (0, \frac{1}{2})$

$$\mathbb{P}\left(N_i(t^*_{i,k}) \leq K(t^*_{i,k}) | X_i^{(1:k-1)}\right) > 1-\varepsilon.$$

□

2) Proof of Lemma 2:

Lemma 2 (Optimality of Generated Observation Times)**.** *For all $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$ at iteration $k \in \mathbb{N}$, an optimal observation time for each station with respect to optimization problem 1 presented in Sec. III is given by*

$$t^*_{i,k} := \frac{N_{max}}{\hat{\lambda}_{i,k}} = N_{max} \frac{\beta_i}{\alpha_i}$$

where $N_{max} := \max_{i \in [n]} \hat{\lambda}_{i,k} t^{low}_{i,k}$ and $t^{low}_{i,k}$ is given by expression (9).

Proof. We argue by contradiction, suppose that there exists some $t^*_{i,k}$ that happens to not be the optimal solution to problem 1. This implies that $t^*_{i,k}$ either (i) violates the uncertainty constraint (3) or (ii) induces an unbalanced observation scheme.

We immediately see that case (i) leads to a contradiction since $t^*_{i,k}$ is defined to be bounded below by the solution to given by the expression in Eq. (9), $t^{\text{low}}_{i,k}$, hence by monotonicity of the uncertainty condition, any value greater than or equal to also satisfies the inequality given by (3). Similarly, we note that (ii) also leads to a contradiction and thus cannot occur since by definition of each $t^*_{i,k}$, we have:

$$\hat{\lambda}_{1,k} t^*_{1,k} = N_{\max}, \hat{\lambda}_{2,k} t^*_{2,k} = N_{\max}, \ldots, \hat{\lambda}_{n,k} t^*_{n,k} = N_{\max}.$$

which implies that $\pi^*_k = (t^*_{1,k}, \ldots, t^*_{n,k})$ maximizes balance (i.e., objective function 2)

$$\begin{aligned}
&\mathbb{E}[N_1(\pi^*_k)] = \mathbb{E}[N_2(\pi^*_k)] = \cdots = \mathbb{E}[N_n(\pi^*_k)] \\
&\iff \pi^*_k \in \operatorname*{argmax}_{\pi_k} f_{\text{bal}}(\pi_k)
\end{aligned}$$

hence, we have that (ii) leads to a contradiction. Since we have exhausted all the cases of sub-optimality, it must be the case that for all stations $i \in [n]$ and all iterations $k \in \mathbb{N}$, the value of $t^*_{i,k}$ is optimal, implying that the policy $\pi^*_k = (t^*_{1,k}, \ldots, t^*_{n,k})$ with respect to the per-cycle optimization problem. □

3) Proof of Lemma 3:

Lemma 3 (Bound on Posterior Variance)**.** *For any $\varepsilon \in (0, \frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, the posterior variance $Var(\lambda_i | X^{(1:k)})$ is bounded above by $\delta^k Var(\lambda_i)$ with probability at least $(1-\varepsilon)^k$, i.e.,*

$$\mathbb{P}\left(Var(\lambda_i | X_i^{(1:k)}) \leq \delta^k Var(\lambda_i) | X^{(1:k)}\right) > (1-\varepsilon)^k$$

for all stations $i \in [n]$ where $Var(\lambda_i) := \frac{\alpha_{i,0}}{\beta_{i,0}^2}$ is the prior variance.

Proof. From Lemma 1 we have that each $t_{i,k}^*$ is ensured to satisfy the uncertainty condition (3) $\forall i \in [n]$

$$\mathbb{P}\left(Var(\lambda_i|X^{(1:k)}) \leq \delta Var(\lambda_i|X_i^{(1:k-1)})|X_i^{(1:k-1)}\right) > 1-\varepsilon \tag{14}$$

for each iteration k regardless of the events that transpire in the other iterations. Hence, the probability of satisfying this condition for k consecutive iterations is greater than $(1-\varepsilon)^k$. This implies that, with probability at least $(1-\varepsilon)^k$, we have that the following chain of of inequalities holds:

$$\begin{aligned}
&Var\big(\lambda_i|X_i^{(1)}\big) \leq \delta Var\big(\lambda_i\big),\\
&Var\big(\lambda_i|X_i^{(1:2)}\big) \leq \delta Var\big(\lambda_i|X_i^{(1)}\big) = \delta^2 Var\big(\lambda_i\big),\\
&\vdots\\
&Var\big(\lambda_i|X_i^{(1:k)}\big) \leq \delta Var\big(\lambda_i|X_i^{(1:k-1)}\big) = \delta^k Var\big(\lambda_i\big)
\end{aligned}$$

□

4) Proof of Corollary 4:

Corollary 4 (Bound on Approximation Variance)**.** *For any $\varepsilon \in (0,\frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, the variance of our approximation $Var\big(\hat{\lambda}_{i,k}|X^{(1:k-1)}\big)$ is bounded above by $\delta^{k-1}Var(\lambda_i)$ with probability greater than $(1-\varepsilon)^{k-1}$, i.e.,*

$$\mathbb{P}\left(Var(\hat{\lambda}_{i,k}|X^{(1:k-1)}) \leq \delta^{k-1}Var(\lambda_i)|X^{(1:k-1)}\right) > (1-\varepsilon)^{k-1}$$

for all stations $i \in [n]$.

Proof. Employing the law of total conditional variance, we have for each $i \in [n]$

$$\begin{aligned}
&Var(\lambda_i|X_i^{(1:k-1)})\\
&\quad = \mathbb{E}[Var(\lambda_i|X^{(1:k)})] + Var(\mathbb{E}[\lambda_i|X^{(1:k)}]|X^{(1:k-1)})\\
&\quad = \mathbb{E}[Var(\lambda_i|X^{(1:k)})] + Var(\hat{\lambda}_{i,k}|X^{(1:k-1)})\\
&\quad \geq Var(\hat{\lambda}_{i,k}|X^{(1:k-1)})
\end{aligned}$$

Invoking Lemma 3, we have that $Var(\lambda_i|X_i^{(1:k-1)}) \leq \delta^{k-1}Var(\lambda_i)$ with probability greater than $(1-\varepsilon)^{k-1}$. Combining this inequality with the above application of law of total conditional variance yields the result. □

5) Proof of Theorem 5:

Theorem 5 (ξ-Bound on the Approximation Error)**.** *For any $\varepsilon \in (0,\frac{1}{2}), \delta \in (0,1)$, after $k \in \mathbb{N}_+$ iterations, for any $\xi \in \mathbb{R}_+$, our approximation $\hat{\lambda}_{i,k}$ lies within a ball of radius ξ centered at λ_i with probability at least $(1-\varepsilon)^{k-1}(1-\frac{\delta^{k-1}Var(\lambda_i)}{\xi^2})$, i.e.,*

$$\mathbb{P}\left(|\hat{\lambda}_{i,k}-\lambda_i| < \xi\,|X^{(1:k-1)}\right) > (1-\varepsilon)^{k-1}\big(1-\frac{\delta^{k-1}Var(\lambda_i)}{\xi^2}\big)$$

for all $i \in [n]$.

Proof. Note that by Chebyshev's inequality states the following:

$$\mathbb{P}\left(|\hat{\lambda}_{i,k}-\lambda_i| < \xi\,|X^{(1:k-1)}\right) > 1-\frac{Var(\hat{\lambda}_{i,k}|X^{(1:k-1)})}{\xi^2}.$$

In light of Corollary 4, we have that

$$\mathbb{P}\left(Var(\hat{\lambda}_{i,k}|X^{(1:k-1)}) \leq \delta^{k-1}Var(\lambda_i)|X^{(1:k-1)}\right) > (1-\varepsilon)^{k-1}$$

employing this inequality and Chebyshev's inequality yields:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\lambda}_{i,k}-\lambda_i| < \xi\,\big|X^{(1:k-1)}\right)\\
&\quad > (1-\varepsilon)^{k-1}\big(1-\frac{Var(\hat{\lambda}_{i,k}|X^{(1:k-1)})}{\xi^2}\big)\\
&\quad > (1-\varepsilon)^{k-1}\big(1-\frac{\delta^{k-1}Var(\lambda_i)}{\xi^2}\big)
\end{aligned}$$

□

6) Proof of Theorem 6:

Theorem 6 (Δ-Bound on Policy Optimality)**.** *For any $\xi_i \in \mathbb{R}_+$, $i \in [n]$, given that $0 < |\hat{\lambda}_{i,k}-\lambda_i| < \xi_i$ with probability as given in Theorem 5, let $\sigma_{min} := \sum_{i=1}^n (\lambda_i-\xi_i)^{-1}$ and $\sigma_{max} := \sum_{i=1}^n (\lambda_i+\xi_i)^{-1}$. Then, the objective value of our policy π_k^* at iteration k is within a factor of Δ of the ground-truth optimal solution, where $\Delta := \frac{\sigma_{min}}{\sigma_{max}}$ with probability greater than $(1-\varepsilon)^{n(k-1)}\big(1-\frac{\delta^{k-1}Var(\lambda_i)}{\xi^2}\big)^n$.*

Proof. Let $T = \sum_{i=1}^n t_{i,k}^*$ be the total observation time allocated by the generated policy. Then, by the optimality of policy $\pi_k^* = (t_{1,k}^*,\ldots,t_{n,k}^*)$ with respect to the rate approximations, we have the following equalities

$$\hat{\lambda}_{1,k}t_{1,k}^* = N_{\max}, \hat{\lambda}_{2,k}t_{2,k}^* = N_{\max}, \ldots, \hat{\lambda}_{n,k}t_{n,k}^* = N_{\max}.$$

which implies that

$$\forall i \in [n] \quad t_{i,k}^* := \frac{T}{\hat{\lambda}_{i,k}\sum_{l=1}^n \frac{1}{\hat{\lambda}_{l,k}}}.$$

Now recall that the objective function pertaining to balance (2) is given by:

$$f_{\text{bal}}(\pi_k) := \min_i \frac{\mathbb{E}[N_i(\pi_k)]}{\sum_{j=1}^n \mathbb{E}[N_j(\pi_k)]}.$$

and the optimal (maximal) value of this function is $\frac{1}{n}$. Now, using the fact that $|\hat{\lambda}_{i,k}-\lambda_i| < \xi_i$, we have the following

inequalities for π_k^*

$$
\begin{aligned}
f_{\text{bal}}(\pi_k^*) &= \min_i \frac{\mathbb{E}[N_i(\pi_k^*)]}{\sum_{j=1}^n \mathbb{E}[N_j(\pi_k^*)]} \\
&= \frac{\min_i \hat{\lambda}_{i,k} t_{i,k}^*}{\sum_{j=1}^n \hat{\lambda}_{j,k} t_{j,k}^*} \\
&= \frac{\min_i \frac{T}{\sum_{l=1}^n (\hat{\lambda}_{l,k})^{-1}}}{\sum_{j=1}^n \frac{T}{\sum_{l=1}^n (\hat{\lambda}_{l,k})^{-1}}} \\
&> \frac{\frac{T}{\sum_{l=1}^n (\lambda_i + \xi_l)^{-1}}}{\frac{nT}{\sum_{l=1}^n (\lambda_l - \xi_l)^{-1}}} \\
&= \frac{\sum_{l=1}^n (\lambda_l - \xi_l)^{-1}}{n \sum_{l=1}^n (\lambda_l + \xi_l)^{-1}} \\
&= \frac{1}{n} \left(\frac{\sigma_{\min}}{\sigma_{\max}} \right)
\end{aligned}
$$

with probability at least $(1-\varepsilon)^{n(k-1)} \left(1 - \frac{\delta^{k-1} Var(\lambda_i)}{\xi^2}\right)^n$. $\square$